UCRL-ID-146978

# Final Report LDRD 99-ERI-010 Sapphire: Scalable Pattern Recognition for Large-Scale Scientific Data Mining

*C. Kamath*

**January 30, 2002**

*U.S. Department of Energy*

Lawrence
Livermore
National
Laboratory

## DISCLAIMER



This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

Final Report
LDRD 99-ERI-010
Sapphire: Scalable Pattern Recognition for Large-Scale Scientific Data Mining

Chandrika Kamath, PI.
Center for Applied Scientific Computing

There is a rapidly widening gap between our ability to collect data and our ability to explore, analyze, and understand the data. As a result, useful information is overlooked, and the potential benefits of increased computational and data gathering capabilities only partially realized. This problem of data overload is becoming a serious impediment to scientific advancement in areas as diverse as counter-proliferation, the Accelerated Strategic Computing Initiative (ASCI), astrophysics, computer security, and climate modeling, where vast amounts of data are collected through observations or simulations. To improve the way in which scientists extract useful information from their data, we are developing a new generation of tools and techniques based on data mining.

Data mining is the semi-automated discovery of patterns, associations, anomalies, and statistically significant structures in data. It consists of two steps – in data pre-processing, we extract high-level features from the data, and in pattern recognition, we use the features to identify and characterize patterns in the data. In this project, our focus is on developing scalable algorithms for the pattern recognition task of classification. Our goal is to improve the performance of these algorithms, without sacrificing accuracy. We are demonstrating these techniques using an astronomy application, namely the detection of radio-emitting galaxies with a bent-double morphology in the FIRST survey. Our research has been incorporated into software to make it easily accessible to LLNL scientists. I next describe our accomplishments in each of these three areas.

**Improving the performance of decision tree software:** Decision tree algorithms are a category of pattern recognition algorithms in which we build a model in the form of a tree structure to discriminate among different objects. The decision tree software takes as input, a list of objects, the features associated with them, and the class label indicating the type of object. This list is referred to as the training set. The decision tree algorithm then creates a tree structure, where each node is a simple decision on a feature. When this tree is applied to a new object with its associated features, but without a label, it results in a class label indicating the type of the new object. Our research in improving the accuracy and scalability of decision tree algorithms has focused on the following:

- Improving oblique trees using evolutionary algorithms: In an oblique tree, a decision at a node uses a linear combination of the features, instead of a single feature. As this is essentially a search in a high dimensional space, we investigated the use of evolutionary algorithms to solve the optimization problem . Our research showed that combining evolutionary algorithms with decision trees resulted in better and faster classifiers. On a data set with 50 features or dimensions, one of our new algorithms, Oblique-ES, was more accurate (79 percent vs. 73 percent) and four times faster than the current best oblique classifier. Another algorithm, Oblique-GA,

which gave the most accurate results (85 percent), was twice as fast. In contrast, the traditional axis-parallel tree, while fast, resulted in accuracy (58 percent) that was just a bit better than making a random decision.

- Improving ensembles of trees by sampling at a node: In ensembles of trees, several different trees are created using the training set and the results are combined through simple voting. In the new algorithm we invented, we create the ensemble by randomizing the decision at each node of the tree using a random sample of the instances for each feature. Using public-domain data sets, we showed that this new technique was more accurate than a single tree and competitive in accuracy with other techniques for creating ensembles by a factor of two to six. Infact, an ensemble of ten trees could be created in less time than it took to create a single tree

- Improving ensembles of trees using histograms: In this second new algorithm, the decision at a node of a tree is randomized by first using histograms to find a best split and then randomly selecting a split point in an interval around the best bin boundary of the histogram. This technique was also more accurate than a single tree and competitive in accuracy with other techniques for creating ensembles, but faster.

**Identification of bent-double galaxies in the FIRST survey:** We conducted our experiments with decision tree algorithms in the context of the classification of radio-emitting galaxies with a bent-double morphology in the FIRST survey. The Faint Images of the Radio Sky at Twenty-cm (FIRST) is an astronomical survey using the Very Large Array at the National Radio Astronomy Observatory (http://sundog.stsci.edu). The FIRST astronomers are surveying more than 10,000 square degrees of the sky, to a flux density limit of 1.0 mJy (milli-Jansky). With the data collected through 1999, FIRST has covered about 8,000 square degrees, producing more than 32,000 two-million pixel images. At a threshold of 1 mJy, there are approximately 90 radio-emitting galaxies, or radio sources, in a typical square degree. The final survey will have almost a million galaxies.

While radio sources exhibit a wide range of morphological types, the FIRST astronomers are particularly interested in galaxies with a bent-double morphology, as they indicate the presence of clusters of galaxies. Figure 1 has two images that were identified manually by scientists as bent-double galaxies. This visual inspection of the radio images, besides being very subjective, has also become increasingly infeasible as the survey has grown in size.

The data from the FIRST survey is available as image maps and a catalog. In Figure 1, we show an image map and the three catalog entries corresponding to one of the bent-doubles present in the image map. These large image maps are mostly ``empty'', that is, composed of background noise that appear as streaks in the image. The FIRST catalog is obtained by processing an image map to fit two-dimensional elliptic Gaussians to each radio source. Each entry in the catalog corresponds to the information on a single Gaussian. This includes, among other things, the coordinates for the center of the

Gaussian, the major and minor axes, the peak flux, and the position angle of the major axis (degrees counter-clock-wise from North).

Our approach to mining the FIRST data for bent-doubles was as follows. We focused on the catalog data as it was easy to work with and a good representation of all but the most complex of radio-emitting galaxies. We first grouped the catalog entries that were close to each other, and then focused on those groups that consisted of two or three catalog entries. This was based on the observation that a single entry galaxy was unlikely to be a bent-double, while four or more entries in a galaxy would make it complex enough to be of interest to astronomers. We next extracted a separate set of features for the two- and three-entry galaxies, focusing on features such as relative distances and angles between entries, that were likely to be robust and invariant to rotation, scaling, and translation. Separating the two- and three-entry galaxies enabled us to have uniform length feature vectors for each. However, it also meant that a small training set (313 examples) was split further into smaller training sets of 118 examples for two-entry and 195 examples for three-entry sources, respectively. Our focus in this work is on the three-entry sources, with the training set consisting of 167 bent-doubles and 28 non-bent-doubles. Note that the training set is unbalanced, with far more bent-doubles than non-bent-doubles.

Once we had extracted an initial set of features, we continued refining the features until the cross-validation error for a decision tree classifier was reduced to about 10%, a number the astronomers felt was sufficient for their use. The tree created using this set of features was then used to classify unlabeled galaxies. Several of these galaxies were shown to the astronomers for validation. Since we wanted to use this new set of galaxies to enhance our training set, we selected a higher percentage of galaxies that had been classified as non-bents. This process of validation is rather tedious and has the drawback of being not only subjective, but somewhat inconsistent as the labels assigned by an astronomer are subject to the drift common to human labelers. Therefore, we were able to validate only 290 galaxies, of which 92 were bents and 198 non-bents.

We performed several experiments with this new validated set of examples. We first combined this set with the original training set of 195 instances, and used the combined set of 485 instances to evaluate the algorithms using cross validation. A single decision trees created with the new training set increased the cross-validation error to 20%. The best results of 17% error were obtained with our new techniques for creating ensembles of decision trees. A closer inspection of the misclassified instances indicated that they belonged to the border-line cases, whose labels were often found to be subjective and inconsistent. Therefore, we went back to the original 195 training set, and proceeded along two directions. First, we tried to reduce the number of features using techniques such as principal component analysis and exploratory data analysis. Next, in addition to decision trees, we tried generalized linear models. The six models created using different features and classifiers were then used to classify the unlabeled galaxies. If all six models agreed with a label, we considered the galaxy to be a bent or a non-bent with high probability. As fewer models agreed with a label, the probability of the galaxy being bent or non-bent was reduced. This approach enabled the astronomers to first focus on those

galaxies with were bent with a high probability. These results were communicated to Prof. Robert Becker, the PI for the FIRST project, for further studies.

Our experiences with the problem of classification of bent-double galaxies led to several observations. The existence of the catalog, the easy access to the data, and the availability of software to read, write, and display the astronomical images, all were an immense help in getting started on the problem. We also found that we had to interact extensively with the astronomers, especially at the beginning, in order to understand the data, how it was collected, identify the features that might be relevant to the classification of bent-doubles, etc. We also learnt that the labels assigned by the astronomers to the galaxies could be inconsistent, esepecially in the border-line cases, which also turned out to be the hardest to classify. This, in addition to the lack of ground truth, made it difficult to obtain a good training set.

**Sapphire software for scalable scientific data mining:** Our research in decision tree algorithms has been incorporated into software to make it accessible to LLNL scientists. During the three year LDRD, we had regular releases of our software, mainly for internal use by our collaborators. The software is written in C++ using object-oriented technology. It is designed to incorporate parallelism, though the current version is serial. In addition to a single decision tree, the software also includes the recent algorithms we developed for ensembles, an extensive library of evolutionary algorithms, several different ways of creating oblique decision trees, several pruning options, splitting criteria and split finders.

In addition to the above activities, we have been professionally active; our work has included:

- Publication of several papers in conferences, workshops and journals
- Organization of workshops and programs such as the series of workshops on Mining Scientific Datasets and the week-long program on Mathematical Challenges in Scientific Data Mining at the Institute for Pure and Applied Mathematics at UCLA
- Participation in program committees for several key conferences in data mining an related topics
- Giving tutorials at data mining conferences
- Interactions with university collaborators, including Prof. Leo Breiman from UC Berkeley, Prof. Padhraic Smyth from UC Irvine, Prof. Manjunath from UC Santa Barabara, Prof. Vipin Kumar from University of Minnesota, and Prof. Bob Becker from UC Davis, who was also our collaborator on the work done with the FIRST survey.
- Mentoring of students, including Matt Giamporcaro From Boston University, Rachel Karchin from UC Santa Cruz, Imelda Kirby from University of Washington Seattle, David Nault from University of Cincinnati, David Littau from University of Minnesota, and Ty Jones from University of Nevada, Reno.
- Filing of several records of inventions and patent applications.

More details on the above are available in the attached list of:

- Publications

- Tutorials
- Book Chapters
- Posters and presentations
- List of patents and records of invention

## Publications

1. C. Kamath, Proceedings of the Fourth Workshop on Mining Scientific Datasets (ed.), August 2001, KDD2001, UCRL-ID-144763. Non-refereed report. No abstract available.

2. Fodor, I. K. and C. Kamath, "Dimension reduction techniques and the Classification of Bent Double Galaxies", Computational Statistics and Data Analysis journal, September 2001. UCRL-JC-144209. In press. Refereed publication.

   **Abstract:** As data mining gains acceptance in the analysis of massive data sets, it is becoming clear that we need algorithms that can handle not only the massive size, but also the high dimensionality of the data. When the number of features or attributes reaches hundreds or even thousands, the computational time for the pattern recognition algorithms can become prohibitive. A common solution to this problem is to reduce the dimensionality, either in conjunction with the pattern recognition algorithm or independent of it. In this paper, we describe how we use techniques from statistics and exploratory data analysis to address the problem of high dimensionality by selecting the features that are relevant to the problem. We discuss our work in the context of an astronomy problem, namely the classification of radio-emitting galaxies with a bent-double morphology using decision trees and generalized linear models. We show that a careful extraction and selection of features is necessary for the successful application of data mining techniques.

3. Fodor, I. K. and C. Kamath, "Denoising through Wavelet Shrinkage: An Empirical Study", submitted for publication to Journal of Electronic Imaging, July 2001. UCRL-JC-144258.

   **Abstract:** Techniques based on thresholding of wavelet coefficients are gaining popularity for denoising data. The idea is to transform the data into the wavelet basis, where the "large" coefficients are mainly the signal, and the "smaller" ones represent the noise. By suitably modifying these coefficients, the noise can be removed from the data. In this paper, we evaluate several two-dimensional denoising procedures using test images corrupted with additive Gaussian noise. We consider global, level-dependent, and subband-dependent implementations of these techniques. Our results, using the mean squared error as a measure of the quality of denoising, show that the SureShrink and the BayesShrink methods consistently outperform the other wavelet-based techniques. In contrast, we found that a combination of simple spatial filters led to images that were grainer with smoother edges, though the error was smaller than in the wavelet based methods.

4. E. Cantú-Paz and C. Kamath, "Inducing Oblique Decision Trees with Evolutionary Algorithms", submitted for publication to Journal of AI Research, UCRL-JC-143718.

**Abstract:** This paper illustrates the application of evolutionary algorithms (EAs) to the problem of oblique decision tree induction. The objectives are to demonstrate that EAs can find classifiers whose accuracy is competitive with other oblique tree construction methods, and that, at least in some cases, this can be accomplished in a shorter time. We performed experiments with a (1+1) evolution strategy and a simple genetic algorithm on public domain and artificial data sets, and we compared the results with three other oblique and one axis-parallel decision tree algorithms. The empirical results suggest that the EAs quickly find competitive classifiers, and that EAs scale up better than traditional methods to the dimensionality of the domain and the number of instances used in training. In addition, we show that the classification accuracy improves when the trees obtained with the EAs are combined in ensembles, and that sometimes it is possible to build the ensemble of evolutionary trees in less time than a single traditional oblique tree.

5. Cantú-Paz, E., "Supervised and Unsupervised Discretization methods for Evolutionary Algorithms", Proceedings GECCO 2001 workshop on Optimization by Building and Using Probabilisitic Models. UCRL-JC-142243. Refereed publication.

   **Abstract:** This paper introduces simple model-building evolutionary algorithms (EAs) that operate on continuous domains. The algorithms are based on supervised and unsupervised discretization methods that have been used as preprocessing steps in machine learning. The basic idea is to discretize the continuous variables and use the discretization as a simple model of the solutions under consideration. The model is then used to generate new solutions directly, instead of using the usual operators based on sexual recombination and mutation. The algorithms are tested with several functions and the results suggest that combining discretizers with EAs may be an interesting path for future developments.

6. Kamath, C., E. Cantu-Paz, I. K. Fodor, N. Tang, "Using data mining to find bent-double galaxies in the FIRST survey", Proceedings, Astronomical Data Analysis, Volume 4477, pp. 11-19, SPIE Annual Meeting, San Diego, July-August 2001. UCRL-JC-143458. Refereed publication.

   **Abstract:** In this paper, we describe the use of data mining techniques to search for radio-emitting galaxies with a bent-double morphology. In the past, astronomers from the FIRST (Faint Images of the Radio Sky at Twenty-cm) survey identified these galaxies through visual inspection. This was not only subjective but also tedious as the on-going survey now covers 8000 square degrees, with each square degree containing about 90 galaxies. In this paper, we describe how data mining can be used to automate the identification of these galaxies. We discuss the challenges faced in defining meaningful features that represent the shape of a galaxy and our experiences with ensembles of decision trees for the classification of bent-double galaxies.

7. Fodor, I. K. and C. Kamath, "On denoising images using wavelet-based statistical techniques", LLNL Technical report, March 2001. UCRL –JC-142357. Non-refereed report.

**Abstract:** Techniques based on thresholding of wavelet coefficients are gaining popularity as approaches to denoising data. The main idea is to transform the data into a different basis, the wavelet basis, where the ``large" coefficients are mainly the signal, and the ``smaller" ones represent the noise. By suitably modifying the coefficients in the new basis, the noise can be removed from the data. Much of the work done in this field has focused on one dimensional data, though a few publications have compared select two-dimensional generalizations. In this paper, we extend several one-dimensional denoising procedures to two dimensions, and provide a comprehensive evaluation of the resulting methods. We show that for images, there are several different ways in which these techniques can be applied. Using test images corrupted by additive Gaussian noise, we compare and contrast the methods across a range of noise levels. Our results, using the mean squared error as a measure of the quality of denoising, show that the SureShrink and the BayesShrink methods consistently outperform the other wavelet-based techniques we considered. We also compare the effectiveness of these methods with simple spatial filters. While no filter was consistently the best, we found that a combination of the minimum mean squared error filter, followed by a Gaussian filter, often led to smaller error than the best wavelet techniques.

8. Kamath, C., "The role of Parallel and Distributed Processing in Data Mining", Spring 2001 newsletter of the IEEE Technical Committee on Distributed Processing, pages 10-15. UCRL-JC-142468. Non-refereed report.

9. Kamath, C. and E. Cantu-Paz, "Creating ensembles of decision trees through sampling", 33-rd Symposium on the Interface of Computing Science and Statistics, June 2001, UCRL-JC-142268. Refereed publication.

   **Abstract:** Recent work in classification indicates that significant improvements in accuracy can be obtained by growing an ensemble of classifiers and having them vote for the most popular class. This paper focuses on ensembles of decision trees that are created with a randomized procedure based on sampling. Randomization can be introduced by using random samples of the training data (as in bagging or boosting) and running a conventional tree-building algorithm, or by randomizing the induction algorithm itself. The objective of this paper is to describe our first experiences with a novel randomized tree induction method that uses a sub-sample of instances at a node to determine the split. Our empirical results show that ensembles generated using this approach yield results that are competitive in accuracy and superior in computational cost to boosting and bagging.

10. Fodor, I. K. and C. Kamath, "A comparison of de-noising techniques for FIRST images", Workshop Proceedings, Third workshop on Mining Scientific Datasets, Chicago, April 2001. pp 13-20. UCRL-JC-142085. Refereed publication.

    **Abstract:** Data obtained through scientific observations are often contaminated by noise and artifacts from various sources. As a result, a first step in mining these data

is to isolate the signal of interest by minimizing the effects of the contaminations. Once the data has been cleaned or de-noised, data mining can proceed as usual. In this paper, we describe our work in de-noising astronomical data from the Faint Images of the Radio Sky at Twenty-Centimeters (FIRST) survey. We are mining this survey to detect radio-emitting galaxies with a bent-double morphology. This task is made difficult by the noise in the images caused by the processing of the sensor data. We compare three different approaches to de-noising: thresholding of wavelet coefficients advocated in the statistics community, traditional filtering methods used in the image processing community, and a simple thresholding scheme proposed by FIRST astronomers. While each approach has its merits and pitfalls, we found that for our purpose, the simple thresholding scheme worked relatively well for the FIRST data set.

11. Kamath, C. "Mining Data for Gems of Information", Research Highlight, Science and Technology Review, September 2000, pages 20-22. UCRL-52000-00-9. Non-refereed report.

12. Kamath, C. and E. Cantú-Paz, On the Design of a Parallel Object-Oriented Data Mining Toolkit, KDD-2000 Workshop on Distributed and Parallel Knowledge Discovery, at KDD-2000, Boston, August, 2000. UCRL-JC-138973. Refereed publication.

    **Abstract:** As data mining techniques are applied to ever larger data sets, it is becoming clear that parallel processors will play an important role in reducing the turn-around time for data analysis. In this paper, we describe the design of a parallel object-oriented toolkit for mining scientific data sets. After a brief discussion of our design goals, we describe our overall system design that uses data mining to find useful information in raw data in an iterative and interactive manner. Using decision trees as an example, we illustrate how the need to support flexibility and extensibility can make the parallel implementation of our algorithms very challenging. We describe the solution approaches we are considering to address these challenges. As this is work in progress, we also present some preliminary results using an astronomy data set.

13. Cantú-Paz, E., On the Effects of Migration on the Fitness Distribution of Parallel Evolutionary Algorithms, Workshop on Evolutionary Computation and Parallel Processing, in GECCO 2000, Las Vegas, NV, July 2000. UCRL-JC-138729. Refereed publication.

    **Abstract:** Migration of individuals between populations may increase the selection pressure. This has the desirable consequence of speeding up convergence, but it may result in an excessively rapid loss of variation that may cause the search to fail. This paper describes the effects of migration on the distribution of fitness. The calculations consider finite populations, arbitrary migration rates, and topologies with different numbers of neighbors. The paper shows that even if different algorithms are configured to produce the same selection intensity, they change the composition of

the population in different ways. The results suggest that migration preserves more diversity as the number of neighbors increases.

14. Cantú-Paz, E., Comparing Selection Methods of Evolutionary Algorithms using the Distribution of Fitness, Information Processing Letters. Available as Lawrence Livermore National Laboratory technical report UCRL-JC-138582, April 2000. In press. Refereed publication.

    **Abstract:** Selection methods are essential components of evolutionary algorithms. To have a better chance of success, the selection method must be balanced with the operators that create new individuals, and therefore, understanding the effect of selection on the population is critical for the development of robust problem solvers. This paper describes several popular selection algorithms and shows how to use the order statistics of the fitness distribution to examine them. The calculations show important differences in the selection methods, even when they are configured to have the same selection intensity.

15. Cantú-Paz, E. and C. Kamath, Combining evolutionary algorithms with oblique decision trees to detect bent double galaxies, presented at the SPIE Annual Meeting, San Diego, July-August, 2000, in Proceedings of SPIE, Volume 4120, pages 63-71. UCRL-JC-138979. Refereed publication.

    **Abstract:** Decision trees have long been popular in classification as they use simple and easy-to-understand tests at each node. Most variants of decision trees test a single attribute at a node, leading to axis-parallel trees, where the test results in a hyperplane which is parallel to one of the dimensions in the attribute space. These trees can be rather large and inaccurate in cases where the concept to be learned is best approximated by oblique hyperplanes. In such cases, it may be more appropriate to use an oblique decision tree, where the decision at each node is a linear combination of the attributes. Oblique decision trees have not gained wide popularity in part due to the complexity of constructing good oblique splits and the tendency of existing splitting algorithms to get stuck in local minima. Several alternatives have been proposed to handle these problems including randomization in conjunction with deterministic hill-climbing and the use of simulated annealing. In this paper, we use evolutionary algorithms (EAs) to determine the split. EAs are well suited for this problem because of their global search properties, their tolerance to noisy fitness evaluations, and their scalability to large dimensional search spaces. We demonstrate our technique on a synthetic data set, and then we apply it to a practical problem from astronomy, namely, the classification of galaxies with a bent-double morphology. In addition, we describe our experiences with several split evaluation criteria. Our results suggest that, in some cases, the evolutionary approach is faster and more accurate than existing oblique decision tree algorithms. However, for our astronomical data, the accuracy is not significantly different than the axis-parallel trees.

16. Kamath, C., C. Baldwin, I. K. Fodor, and N.Tang, On the Design and Implementation of a Parallel, Object-Oriented, Image Processing Toolkit, SPIE Annual Meeting, San Diego, July-August 2000, In Proceedings of SPIE, volume 4118. UCRL-JC-138953. Refereed publication.

    **Abstract:** Advances in technology have enabled us to collect data from observations, experiments, and simulations at an ever increasing pace. As these data sets approach the terabyte and petabyte range, scientists are increasingly using semi-automated techniques from data mining and pattern recognition to find useful information in the data. In order for data mining to be successful, the raw data must first be processed into a form suitable for the detection of patterns. When the data is in the form of images, this can involve a substantial amount of processing on very large data sets.

    To help make this task more efficient, we are designing and implementing an object-oriented image processing toolkit that specifically targets massively-parallel, distributed-memory architectures. We first show that it is possible to use object-oriented technology to effectively address the diverse needs of image applications. Next, we describe how we abstract out the similarities in image processing algorithms to enable re-use in our software. We will also discuss the difficulties encountered in parallelizing image algorithms on massively parallel machines as well as the bottlenecks to high performance. We will demonstrate our work using images from an astronomical data set, and illustrate how techniques such as filters and denoising through the thresholding of wavelet coefficients can be applied when a large image is distributed across several processors.

17. Cantú-Paz, E. and C. Kamath, Using Evolutionary Algorithms to Induce Oblique Decision Trees, in Proceedings of the Genetic and Evolutionary Computation Conference (GECCO) 2000, Eds. D. Whitley, D. Goldberg, E. Cantu-Paz, L. Spector, I. Parmee, and H.G.Beyer, pages 1053-1060. UCRL-JC-137202. Refereed publication.

    **Abstract:** This paper illustrates the application of evolutionary algorithms (EAs) to the problem of oblique decision tree induction. The objectives are to demonstrate that EAs can find classifiers whose accuracy is competitive with other oblique tree construction methods, and that at least in some cases this can be accomplished in a shorter time. Experiments were performed with a (1+1) evolution strategy and a simple genetic algorithm on public domain and artificial data sets. The empirical results suggest that the EAs quickly find competitive classifiers, and that EAs scale up better than traditional methods to the dimensionality of the domain and the number of instances used in training.

18. Cantú-Paz, E. Selection Intensity in Genetic Algorithms with Generation Gaps, in Proceedings of the Genetic and Evolutionary Computation Conference (GECCO) 2000, Eds. D. Whitley, D. Goldberg, E. Cantu-Paz, L. Spector, I. Parmee, and H.G.Beyer, pages 911-918. UCRL-JC-137169. Refereed publication.

**Abstract:** This paper presents calculations of the selection intensity of common selection and replacement methods used in genetic algorithms (GAs) with generation gaps. The selection intensity measures the increase of the average fitness of the population after selection, and it can be used to predict the number of steps until the population converges to a unique solution. The theory may help to explain the fast convergence of some algorithms with small generation gaps. The accuracy of the calculations was verified experimentally with a simple test function. The results facilitate comparisons between different algorithms, and provide a tool to adjust the selection pressure, which is indispensable to obtain robust algorithms.

19. Fodor, I. K., E. Cantú-Paz,, C. Kamath, and N. Tang, Finding Bent-Double Radio Galaxies: A Case Study in Data Mining, Interface: Computer Science and Statistics, Volume 32, New Orleans, LA, April 2000, pp 37-47. UCRL-JC-138073 (article), and UCRL-JC-138073 (viewgraphs). Refereed publication.

**Abstract:** This paper presents our early results in applying data mining techniques to the problem of finding radio-emitting galaxies with a bent-double morphology. In the past, astronomers on the FIRST (Faint Images of the Radio Sky at Twenty-cm) survey have detected such galaxies by first inspecting the radio images visually to identify probable bent-doubles, and then conducting observations to confirm that the galaxy is indeed a bent-double. Our goal is to replace this visual inspection by a semi-automated approach. In this paper, we present a brief overview of data mining, describe the features we use to discriminate bent-doubles from non-bent-doubles, and discuss the challenges faced in defining meaningful features in a robust manner. Our experiments show that data mining, using decision trees, can indeed be a viable alternative to the visual identification of bent-double galaxies.

20. Cantú-Paz E., Genetic Algorithms, Encyclopedia of Computers and Computer History, Chicago, IL: Fitzroy Dearborn, UCRL-JC-137552. Non-refereed report. No abstract available.

## BOOK CHAPTERS

1. Kamath, C., and R. Musick, Scalable Data Mining through Fine-Grained Parallelism: The Present and the Future, chapter in Advances in Distributed and Parallel Knowledge Discovery, H. Kargupta and P. Chan, Eds., AAAI Press, 2000, pp 29-77. UCRL-JC-133694. Refereed publication.

2. Kargupta, H., C. Kamath, and P. Chan, Distributed and Parallel Data Mining: Emergence, Growth, and Future Directions, concluding chapter in the book, Advances in Distributed and Parallel Knowledge Discovery, H. Kargupta and P. Chan, Eds., AAAI Press, 2000, pp 409-417. UCRL-JC-138954. Refereed publication.

3. C. Kamath, "On Mining Scientific Datasets", in Data Mining for Scientific and Engineering Applications, eds. R. Grossman, C. Kamath, W. Kegelmeyer, V. Kumar, and R. Namburu, Kluwer, pp. 1-22, 2001. UCRL-JC-141709. Refereed publication.

4. Kamath, C, E. Cantu-Paz, I. K. Fodor, N. Tang, "Searching for Bent-Double Galaxies in the FIRST Survey", in Data Mining for Scientific and Engineering Applications, eds. R. Grossman, C. Kamath, W. Kegelmeyer, V. Kumar, and R. Namburu, Kluwer, pp. 95-114, 2001. UCRL-JC-140418. Refereed publication.

5. Fodor, I. K. and C. Kamath, "The Role of Multiresolution in Mining Massive Image Datasets", in Multiscale and Multiresolution Methods, Lecture Notes in Computational Science and Engineering, T.J. Barth, T. Chan, and R. Haimes (eds.), Springer-Verlag, November 2001, Volume 20, pages 307-318. UCRL-JC-139713. Refereed publication.

6. Cantu-Paz, E. and C. Kamath, "On the Use of Evolutionary Algorithms in Data Mining", in Data Mining: A Heuristic Approach, Eds. H. Abbass, R. Sarker, and C. Newton., pages 48-71. UCRL-JC-141872. Refereed publication.

## TUTORIALS

1. Kamath, C., "Data Mining for Science and Engineering Applications", Tutorial at the First SIAM conference on data mining, Chicago, April 5-7, 2001. UCRL-JC-142626.

2. Cantú-Paz, E., "Parallel Genetic Algorithms", tutorial at GECCO, 2001. UCRL-VG-136736

3. R. Grossman, C. Kamath, V. Kumar, "Data mining for scientific and engineering applications", Tutorial at Supercomputing 2001, November 2001. UCRL-PRES-145087.

## POSTERS, PRESENTATIONS, AND OTHER PUBLICATIONS

This list includes the posters and presentations that did not require a paper. If a presentation was done at a conference or workshop, with an attached paper, it is listed under the papers. This list also includes web pages, and brochures.

1. Becker, Bob, and Chandrika Kamath, "Statistics, Pattern Recognition, and Astrophysics," Neyman seminar conducted for the University of California, Berkeley, Department of Statistics, October 14, 1998. UCRL-MI-132093

2. Kamath, Chandrika, "Sapphire: Data Mining and Pattern Recognition for Large and Complex Science Data," Women's Technical and Professional Symposium, San Ramon, CA, October 15-16, 1998. UCRL-MI-132151.

3. Kamath, C. Sapphire Large Scale Data Mining and Pattern Recognition, Web pages located at http://www.llnl.gov/casc/sapphire, UCRL-MI-1341567.

4. Kamath, C. SAPPHIRE: Large-Scale Data Mining and Pattern Recognition, LLNL Technical brochure, UCRL-TB-132076, January 1999.

5. Kamath, C. Sapphire: An Object-oriented Framework for Mining Science Data, presented at the First Workshop on Mining Scientific Datasets, Minneapolis, MN, September 1999. Available as Lawrence Livermore National Laboratory technical report UCRL-JC-135563-abs (abstract), and UCRL-VG-135563 (viewgraphs), September 1999.

6. Baldwin, C. and C. Kamath, Denoising Data Using Wavelet Based Methods, Signal and Imaging Sciences Workshop, Lawrence Livermore National Laboratory, November 11-12, 1999. Available as Lawrence Livermore National Laboratory technical report UCRL-JC-136075-abs (abstract), and UCRL-VG-136075 (viewgraphs), November 1999.

7. Fodor, I. K., Duffy, P., Kamath, C. and Baldwin, C., Effect of Missing Data on the Apparent Trend in the Earths Surface Temperature since 1860, American Geophysical Union Annual Fall Meeting, San Francisco, CA, 1999. Available as Lawrence Livermore National Laboratory technical report UCRL-VG-135586, December 1999.

8. Kamath, C., E. Cantú-Paz, N. Tang, Sapphire: A High-Performance Object-Oriented Framework for Mining Scientific Datasets, Third International Symposium for Computing in Object-oriented Parallel Environments, San Francisco, CA, December 1999. Available as Lawrence Livermore National Laboratory technical report UCRL-JC-136633-abs (abstract), UCRL-VG-136633 (viewgraphs), December 1999.

9. Nault, D., Parallel Object-Oriented File I/O for the Sapphire Project, Poster for the STEP Program, Available as Lawrence Livermore National Laboratory technical report UCRL-MI-136435, December 1999.

10. Kamath, C., "Distributed and Parallel Data Mining: Advances and Future Directions", Panel discussion at the Workshop on Distributed and Parallel Knowledge Discovery, at the Knowledge Discovery and Data Mining Conference, Boston, August 20-23, 2000. Also available as Lawrence Livermore National Laboratory technical report UCRL-VG-140067.

11. Kamath, C., On Oblique Decision Trees and Evolutionary Algorithms, invited presentation at the Second Workshop on Mining Scientific Datasets, July 20-21, Minneapolis, Minnesota, 2000. Available as Lawrence Livermore National Laboratory technical report UCRL-JC-139515 abs.

12. Fodor, I. K., C. Kamath, E. Cantu-Paz, N. Tang, The Search for Bent-double Galaxies: Latest Results, invited presentation at the Second Workshop on Mining Scientific Datasets, July 20-21, Minneapolis, Minnesota, 2000. Available as Lawrence Livermore National Laboratory technical report UCRL-JC-139561 abs.

13. Fodor, I. K. and C. Kamath, "Wavelet-Based Denoising Techniques in Sapphire", presentation at the CASIS Workshop, November 16-17, LLNL, 2000. UCRL-JC-141298 abs, and UCRL-VG-141298.

14. Cantu-Paz, E. and C. Kamath, "Improving the Performance of Linear Decision Trees with Evolutionary Algorithms", presentation at the CASIS Workshop, November 16-17, LLNL, 2000. UCRL-JC-141297 abs, and UCRL-VG-141297.
15. Fodor, I. K., "Statistical Issues in Data Mining", presentation at the Fifth North American Meeting of New Researchers in Statistics and Probability, Georgia Institute of Technology, Atlanta, July 2001. UCRL-JC-142212-abs.

16. Cantu-Paz, E. and C. Kamath, "Sapphire: Mining Scientific Datasets", invited presentation at the Joint Statistical Meetings, August 2001, Atlanta. UCRL-JC-142044-abs.

17. Fodor, I. K. and C. Kamath, "Non-linear methods in data mining: Comparison of 2D Wavelet De-noising Methods", poster at the MSRI workshop on Nonlinear Estimation and Classification, March 2001, Berkeley. UCRL-JC-141738abs

18. Kamath, C., "Mining Science Data: The Sapphire Approach", presentation at the Colloquium, Computer Science and Electrical Engineering Departments, University of Nevada, Reno, April 19, 2001, UCRL-JC-143254 abs.

19. Cantú-Paz, E., "Single vs. Multiple Runs under Constant Computation Cost", poster presentation at GECCO 2001, UCRL-JC-142172.

## PATENTS AND RECORDS OF INVENTION

1. Kamath, C., E. Cantú-Paz, and D. Littau, "Using histograms to introduce randomization in the generation of ensembles of decision trees", Record of Invention, LLNL Invention case number IL 10905, July 2001.

2. Cantú-Paz, E. and C. Kamath, "Creating ensembles of oblique decision trees with evolutionary algorithms and sampling", Record of Invention, LLNL invention case number IL 10877, May 2001.

3. Kamath, C. and E. Cantú-Paz, "Generating ensembles of decision trees by sampling the data instances at each node of the tree", Record of Invention, LLNL invention case number IL 10878, May 2001

4. Kamath, C. and E. Cantú-Paz, "Parallel object-oriented data mining system", Patent application, LLNL case number IL 10713, June 2001.

5. Kamath, C. and E. Cantú-Paz, "Parallel object-oriented decision tree system", Patent application, LLNL case number IL 10714, June 2001.

6. Kamath, C., C. Baldwin, I.K. Fodor, and N. Tang, "Parallel object oriented denoising system using wavelet multiresolution analysis", Patent application, LLNL case number 10716, June 2001.